Deep Analysis Results - 80,000 Reviews, 20 Apps

1. Rating Distribution

Metric	Value
Mean	3.71
Median	5.0
Std Dev	1.71
Skewness	-0.77 (left-skewed)
Kurtosis	-1.23 (light tails, bimodal shape)

The distribution is **strongly bimodal**, a large 5-star peak (58.6%) with a secondary 1-star peak (24.5%). The middle ratings (2-4 stars) are sparse. This is characteristic of app store reviews where users tend to either love or hate the experience.

Sentiment buckets: 66.1% positive (4-5), 4.9% neutral (3), 29.0% negative (1-2). The Pos:Neg ratio is 2.28:1, suggesting there's a meaningful class imbalance that downstream models will need to handle (e.g., stratified sampling, class weights).

2. Text Length - Strong Inverse Correlation with Rating

Star	Median Chars	Mean Chars	% <= 10 chars
5	11	29	47.5%
4	20	60	36.2%
3	51	103	23.6%
2	97	143	10.8%
1	80	135	8.9%

Negative reviews are 5x longer than positive ones. Nearly half of all 5-star reviews are 10 characters or less ("good", "nice"). This creates a structural imbalance, meaning positive reviews carry less textual signal, while negative reviews are information-dense.

Overall, the median review is just 21 characters (4 words). The mean is pulled up to 66 chars by the long tail (P95=315, P99=494, max=500, Google Play truncates at 500 chars).

3. Temporal Patterns

- **Span**: 100 days (Oct 27, 2025 - Feb 4, 2026)
- **Recency skew**: Volume increases exponentially toward the present (12,067 reviews on Feb 3 vs ~30-50/day in October). This is expected with the "newest" sort order.
- **Daily average rating is stable** around 3.5-3.8, no major temporal drift.
- **Hour-of-day**: Peak at midnight-1am UTC (evening US time), trough at 6-8am UTC (early morning US).
- **Day-of-week**: Tuesday peaks, Saturday troughs.

4. Per-App Variance

Apps differ meaningfully on every dimension:

Signal	Low	High
Mean rating	Amazon (2.71), Clash Royale (2.83)	WhatsApp (4.29), Google Photos (4.20)
Avg words	WhatsApp (5.8), Chrome (6.7)	Amazon (27.4), Clash Royale (18.1)
Short review %	Amazon (16.4%)	WhatsApp (66.0%), Chrome (64.0%)
Dev reply rate	0% (9 apps)	TikTok (82.6%), Cash App (63.8%)

Cross-app rating variance is 0.197. Models should be aware that app identity is a strong confounder.

5. Data Quality Issues

Duplicates: 0 duplicate IDs, but 34.3% of rows share content with at least one other row. Dominated by generic phrases: "good" (4,954x), "nice" (1,777x), "Good" (1,335x). These aren't true duplicates, they're independent reviews that happen to say the same thing.

Field completeness: Core fields (review_id, author, rating, content, timestamp) are >99.99% complete. app_version is 85.9% filled, with worst rates on TikTok (31.7% missing), Telegram (29.9%), and Snapchat (26.6%).

Low-signal reviews: 39.1% of all reviews are either empty, single-word, 2-3 words, or non-Latin only. **~61% of the dataset is usable for labeling** at a basic quality threshold.

6. Thumbs-Up (Helpfulness)

82.8% of reviews have zero thumbs-up. Among non-zero: median is 1, mean is 25.7, max is 108,844. Negative reviews receive far more thumbs-up (1-star mean: 14.4) vs positive (5-star mean: 0.75). This metric could serve as a proxy for review informativeness.

7. Developer Replies

Replies correlate strongly with negative sentiment: replied reviews average 2.93 stars vs 3.84 for unreplied (delta: -0.91). This is expected since developers prioritize responding to complaints. Reply behavior varies dramatically by app (0% for 9 apps, 82.6% for TikTok).

Implications for Downstream Work

1. **Filtering needed**: ~39% low-signal reviews should be filtered or flagged before labeling
2. **Class imbalance**: 2.3:1 positive/negative ratio; neutral class is tiny (4.9%)
3. **Text length is a confound**: Positive reviews are sparse in content, sentiment models may need length-aware strategies
4. **Case normalization**: "good" vs "Good" appear as separate entries, text preprocessing should lowercase
5. **App as feature**: Strong per-app differences mean app_id may be a useful feature (or a bias to control for)